Food Quality and Preference 40 (2015) 116–124

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Measures of association between two datasets; Application to sensory data

Angélina El Ghaziri *, El Mostafa Qannari

LUNAM Université, ONIRIS, Unité de sensométrie et Chimiométrie, Nantes F-44322, France
INRA, Nantes, France

ARTICLE INFO

Article history:
Received 23 May 2014
Received in revised form 3 September 2014
Accepted 28 September 2014
Available online 6 October 2014

2010 MSC:
00-01
99-00

Keywords:
Multivariate correlation coefficient
Procrustes similarity index
RV coefficient
Permutation test
Adjusted RV coefficient

ABSTRACT

We review three measures of association between two datasets in view of their use in sensory data. The aim is threefold: (i) to show in which situations each measure of association is appropriate, (ii) to show their properties and how they can be applied efficiently to sensory data, (iii) to compare them. The three measures of association are multivariate correlation coefficient, RV coefficient and Procrustes similarity index. A particular emphasis is put on RV coefficient since it is very popular among sensory scientists. We stress the properties and shortcomings of this coefficient and propose an adjusted RV coefficient to be used instead of RV coefficient, particularly in situations where the number of samples is small or/and the number of variables is large.

Introduction

In sensory evaluation, it often occurs that practitioners are faced with the task of handling two or more datasets. In such a situation, it is very convenient to assess the pairwise 'correlation' or 'similarity' between these datasets. To some extent, this correlation coefficient may be seen as an extension of the correlation coefficient between two variables or some transformation of this coefficient (*e.g.* absolute value, squared value). Ramsay, Berge, and Styan (1984) discuss five 'correlation' coefficients between two datasets among which we find the multivariate correlation coefficient and RV coefficient. Both these coefficients are discussed in this paper together with Procrustes similarity index, which is tightly connected with Generalized Procrustes Analysis (GPA; Gower (1975)).

The multivariate correlation coefficient is a straightforward extension of Pearson correlation coefficient between two variables. It can be applied to two datasets that pertain to the same individuals and the same variables. Typically, this is the case of conventional (or fixed vocabulary) sensory profiling. When using this similarity index, we ideally expect that the two datasets at hand agree with respect to the successive variables (*i.e.* variable j in the first dataset agrees with variable j in the second dataset, with j running from the first to the last variable).

As stated above, Procrustes index is tightly linked to GPA. It is a fact that practitioners of sensory analysis are more familiar with the concept of Procrustes distance between two datasets than to the Procrustes similarity index discussed herein. The connection between these two quantities is very tight and to a very large extent bears resemblance to the connection between the distance between two vectors and the cosine of these two vectors. The interest of working with the similarity index instead of the Procrustes distance is that the former is unit free and insensitive to a multiplication of one or both datasets by a scalar. Moreover, it is much easier to communicate than Procrustes distance as it is bounded by 0 and 1 with very clear guidelines as to how interpret these two bounds. As a by-product of these properties, it follows that, unlike Procrustes distance, Procrustes similarity index can be used to compare different situations although some caution is required. For instance, one can be interested in comparing the agreement among assessors for two different panels.

* Corresponding author at: LUNAM Université, ONIRIS, Unité de sensométrie et Chimiométrie, Nantes F-44322, France.
E-mail address: angelina.el-ghaziri@oniris-nantes.fr (A. El Ghaziri).

http://dx.doi.org/10.1016/j.foodqual.2014.09.010
0950-3293/

Procrustes similarity index was discussed by Sibson (1978). Lorenzo-Seva and ten (2006) also hint to this coefficient and give several references where it is used. In the context of sensory analysis, the interest of this index was discussed by Qannari, Halliday, and Courcoux (1998). It is worth noting that the pairwise Procrustes similarity indices between a set of configurations are given as standard outputs when running the function "GPA()" in the R package FactoMineR (Husson, Josse, Le, & Mazet, 2013).

A particular emphasis is put on RV coefficient because it has become a very popular tool in sensory analysis since its introduction to this field by Schlich (1996). The typical situations in sensory analysis where this coefficient is used are the following:

i. Pairwise comparisons of several configurations associated with various panelists (Mammasse & Schlich, 2014; Vidal et al., 2014). In particular, this is systematically done when running the STATIS method (Abdi, Valentin, Chollet, & Chrea, 2007; Meyners, 2003; Pizarro, Esteban-Díez, Rodríguez-Tecedor, & González-Sáiz, 2013; Schlich, 1996).
ii. Configurations (very often depicted by graphical displays) obtained by means of two methods of evaluation. The aim is to assess to which extent the two methods of evaluation leads to similar outputs (Faye et al., 2006; Reinbach, Giacalone, Ribeiro, Bredie, & Bom Frst, 2014).
iii. Configurations (very often depicted by graphical displays) obtained as outputs of two statistical methods performed on the same data. Obviously, the aim is to assess to which extent the two statistical methods lead to similar outputs (Peltier, Visalli, & Schlich, 2015; Thomsen, Gourrat, Thomas-Danguin, & Guichard, 2014).

The popularity of RV coefficient owes very much to the fact that it enjoys interesting properties, which we shall recall in this paper. However, it also has some limitations that were pointed out by several authors. The main limitation is that it is very sensitive to the number of individuals (*i.e.* rows) and the number of variables (*i.e.* columns) of the datasets at hand (Smilde, Kiers, Bijlsma, Rubingh, & Van Erk, 2009; Tomic, Forde, Delahunty, & Ns, 2013). To counteract this problem Smilde et al., 2009 proposed a modified RV coefficient. We propose herein yet another correction to this problem by introducing a new coefficient called adjusted RV coefficient. On the basis of a simulation study, we show the advantage of this latter coefficient over the modified RV coefficient.

It is of paramount interest to have in mind the properties of the various coefficients because they may be appropriate in some situations and inappropriate in others. For instance, suppose that, within a conventional sensory profiling context, we aim at assessing the agreement between the configurations of two assessors A and B (say). Suppose, for the sake of argument, that for one or more attributes, assessor A gives diametrically opposed assessments of the products than assessor B. In this case, we would expect the 'correlation coefficient' to reflect a high disagreement between A and B. This is indeed the case for the multivariate correlation coefficient, but not for RV coefficient and Procrustes similarity index since these latter coefficients are invariant by rotation and reflection. Furthermore, for two coefficients that basically serve the same purpose, namely RV coefficient and Procrustes similarity index, it is of interest to understand some technical differences between them.

We also discuss hypotheses tests to assess the significance of the various measures of association. More precisely, we discuss permutation tests and, except for Procrustes similarity test, we recall approximations of these tests that are less computationally involving.

The three measures of association are investigated one after the other. For each case, we give properties that shed some light on the interest of the measure of association at hand. We also give illustrations based on real or simulated data.

Multivariate correlation coefficient

Definition and properties

Let us denote by X and Y two datasets measured on the same individuals. For instance, X could be obtained from the sensory evaluation given by an assessor in the course of a conventional sensory profiling, and Y could be the data obtained from yet another assessor or could be the average data set over all the assessors. Throughout this paper all the datasets are assumed to be centered. The multivariate correlation coefficient between X and Y is given by:

$$(MR(X,Y) = \frac{trace(X^TY)}{\sqrt{trace(X^TX)}\sqrt{trace(Y^TY)}}$$

where X^T and Y^T denote the transpose matrices X and Y, respectively and trace is the sum of the main diagonal elements. The expression of this coefficient assumes that X and Y are of the same dimensions (same number of rows and columns). Coefficient MR enjoys the same properties as the Pearson correlation coefficient between two variables, namely, it ranges between -1 and 1. It is equal to -1 (resp., +1) if $X = \alpha Y$ with α a negative (resp., positive) scalar.

In order to see in which situations $MR(X,Y)$ is equal to zero, we can note that the numerator of $MR(X,Y)$ can be expressed as $n\sum_{j=1}^{p} cov(x_j, y_j)$, where n is the number of samples (*i.e.* rows), $cov(x_j, y_j)$ stands for the covariance between variables x_j and y_j. We recall that the covariance is tightly linked to the correlation coefficient since we have for two variables x and y:

$$cov(x,y) = s_x s_y cor(x,y)$$

where s_x and s_y, are respectively, the standard deviation of x and y and $cor(x,y)$ is the Pearson correlation coefficient between x and y. The latter expression of the numerator of $MR(X,Y)$ shows that each variable x_j in X is compared to its counterpart y_j in Y (note the same subscript for x_j and y_j). Therefore, we should expect $MR(X,Y)$ to be zero if the covariances between x_j and y_j are, on average, equal to zero. Roughly speaking, this means that variables x_j and y_j are, on average, uncorrelated. However, this does not mean that each correlation between a variable from X and its corresponding variable in Y is equal or close to 0, but it means that negative covariances between some variables are compensated by positive covariances between other variables.

Another expression of $MR(X,Y)$, which sheds more light on the properties of this coefficient is the following. Suppose that the variables in X (resp., Y) are stacked up vertically, resulting in a single column, which we denote by $vec(X)$ (resp., $vec(Y)$), then $MR(X,Y)$ is simply the Pearson correlation coefficient between vectors $vec(X)$ and $vec(Y)$. In particular, if X and Y are univariate (i.e. $X = [x]$ and $Y = [y]$) then $MR(X,Y)$ boils down to Pearson correlation coefficient between x and y.

Permutation test

Once $MR(X,Y)$ is computed, the question that emerges is to test whether this coefficient is equal to 0, against the alternative hypothesis stipulating that it is larger than 0. In other words we are interested in a Right tailed test:

$$\begin{cases} H_0 : MR(X,Y) = 0 \\ H_1 : MR(X,Y) > 0 \end{cases}$$

In the context of conventional profiling where the aim is to compare the agreement of two configurations associated with